

## Supplementary Information

### **Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genomes**

Haoyang Cai<sup>1,2</sup>, Nitin Kumar<sup>1,2</sup>, Homayoun C. Bagheri<sup>3</sup>, Christian von Mering<sup>1,2</sup>, Mark D. Robinson<sup>1,2</sup>, and Michael Baudis<sup>1,2</sup>

<sup>1</sup>Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

<sup>2</sup>Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

<sup>3</sup>Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

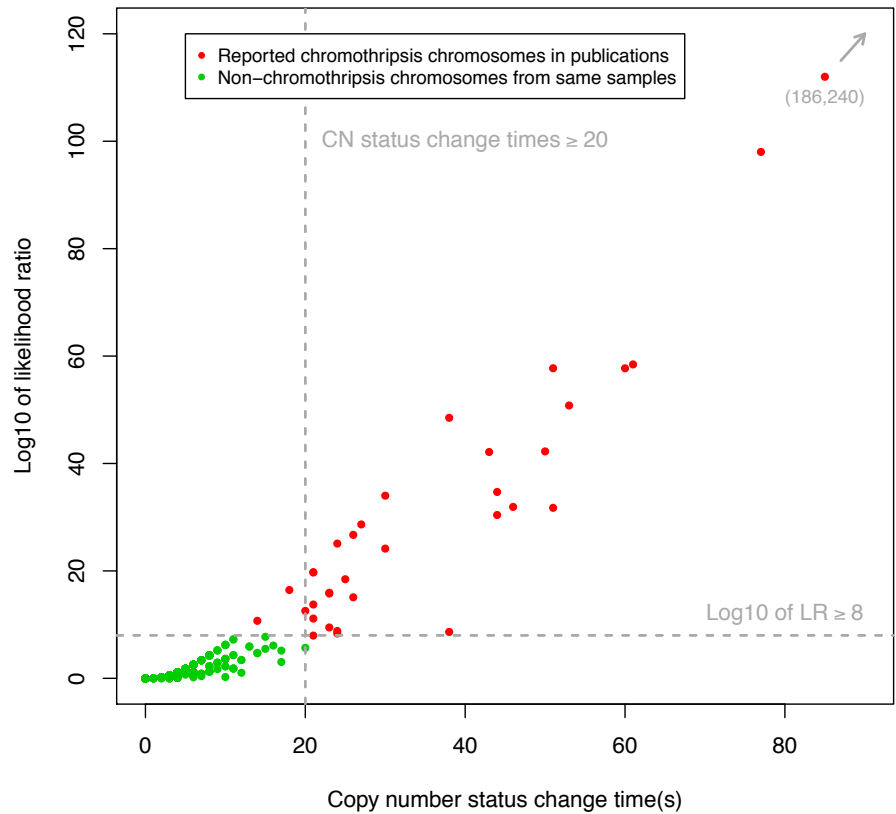
## 1. Supplementary Figures

<b>Supplementary Figure 1</b>	<b>page 3</b>
Scatter plot of the training set.	
<b>Supplementary Figure 2</b>	<b>page 11</b>
The positive training set and CTLP detection algorithm performances.	
<b>Supplementary Figure 3</b>	<b>page 12</b>
Scatter plot of CTLP candidates.	
<b>Supplementary Figure 4</b>	<b>page 13</b>
Kaplan-Meier survival curves for CTLP versus non-CTLP cases in specific cancer types.	
<b>Supplementary Figure 5</b>	<b>page 14</b>
An example of the platform resolution based simulation from Affymetrix SNP6 array (1.8M).	
<b>Supplementary Figure 6</b>	<b>page 15</b>
CTLP detection sensitivity of simulated platform resolutions.	

## 2. Supplementary Tables

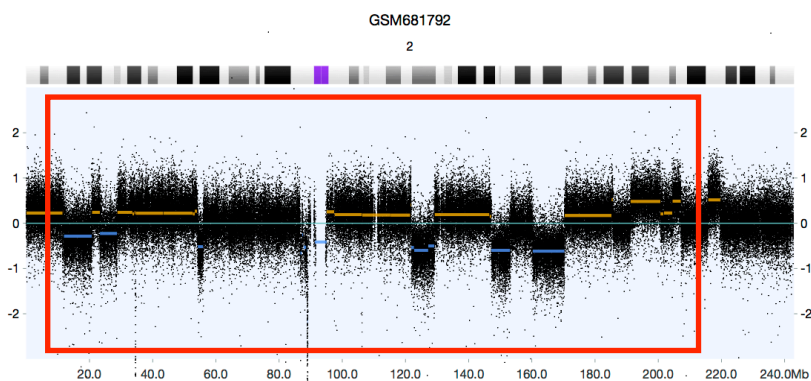
<b>Supplementary Table 1</b>	<b>page 16</b>
Overview of input dataset	
<b>Supplementary Table 7</b>	<b>page 17</b>
Demographic and clinicopathologic characteristics of input and CTLP samples	
<b>Supplementary Table 8</b>	<b>page 18</b>
Sizes of sliding windows for the scan-statistic based algorithm	

Additional File 1: Supplementary Figures and Tables

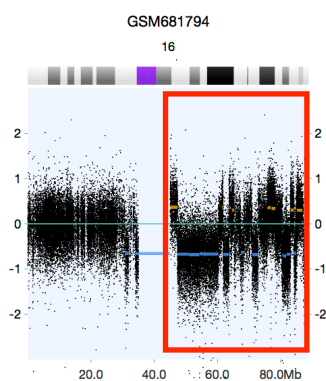


**Figure S1. Scatter plot of the training set.** Copy number status change times compared to the likelihood ratio. Each point represents the window with the highest LR for each chromosome. The dashed lines indicate the selected thresholds. CN, copy number; LR, likelihood ratio.

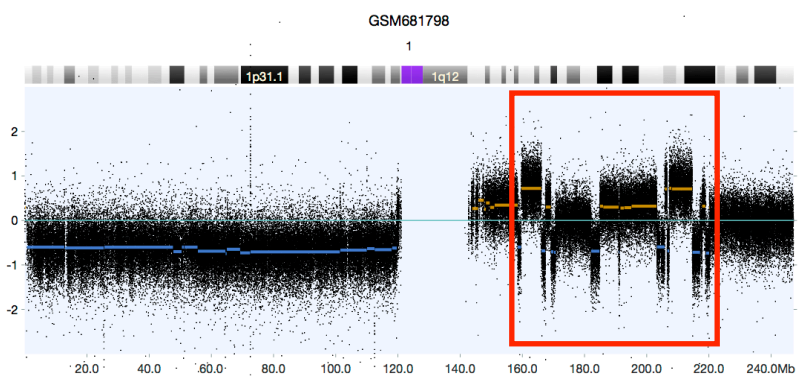
a



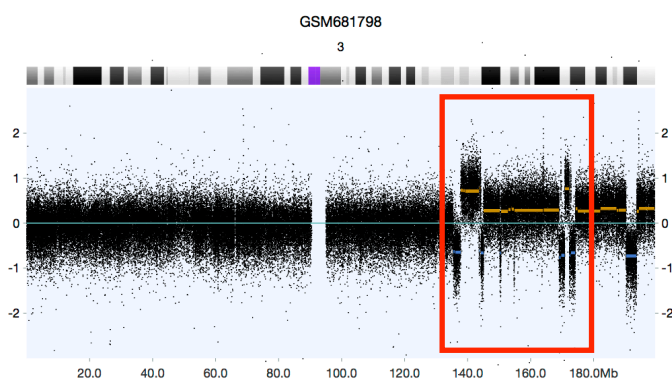
Log10(LR) = 8.6  
Switch times = 38  
Window size = 199.5 Mb



Log10(LR) = 31.9  
Switch times = 46  
Window size = 46.9 Mb



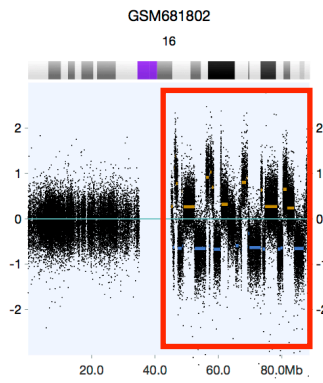
Log10(LR) = 15  
Switch times = 26  
Window size = 62.4 Mb



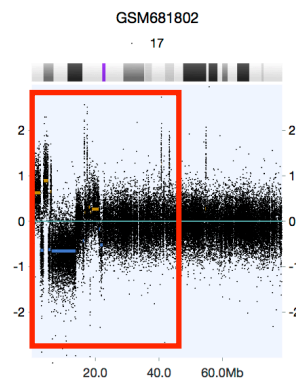
Log10(LR) = 24.2  
Switch times = 30  
Window size = 40 Mb



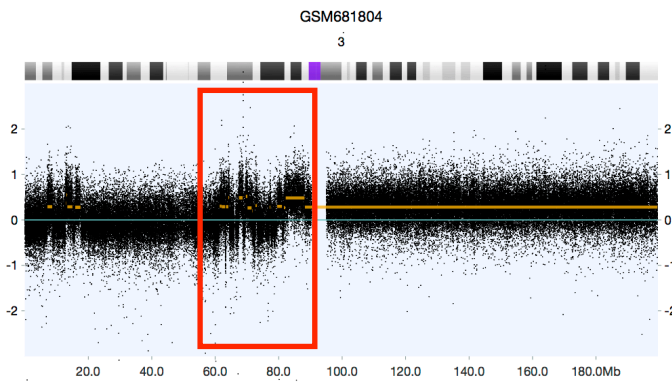
b



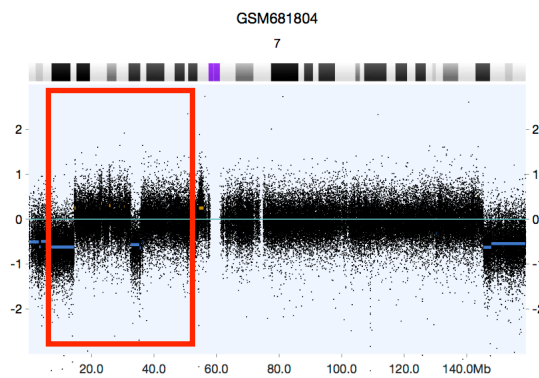
Log10(LR) = 42.1  
Switch times = 43  
Window size = 46.9 Mb



Log10(LR) = 15.9  
Switch times = 23  
Window size = 46.9 Mb

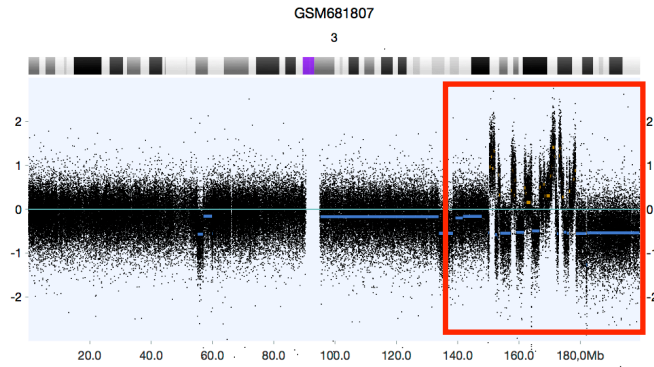


Log10(LR) = 19.7  
Switch times = 21  
Window size = 30 Mb

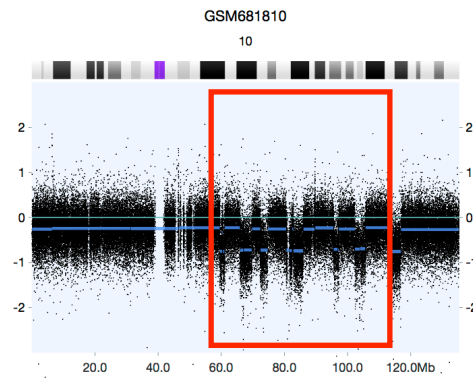


Log10(LR) = 19.7  
Switch times = 21  
Window size = 40 Mb

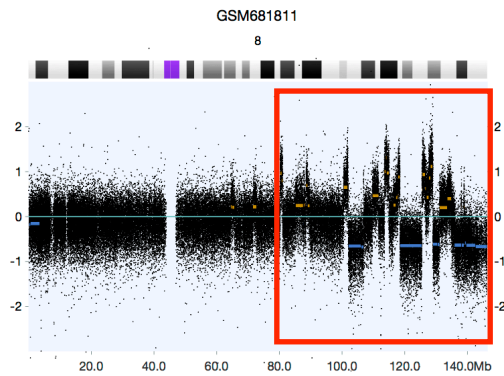
C



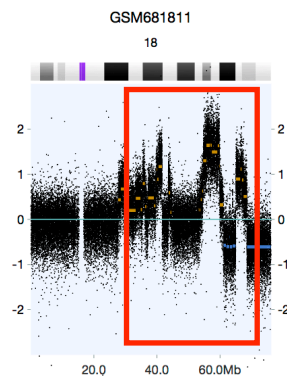
Log10(LR) = 58.4  
Switch times = 61  
Window size = 49.7 Mb



Log10(LR) = 13.7  
Switch times = 21  
Window size = 46.9 Mb

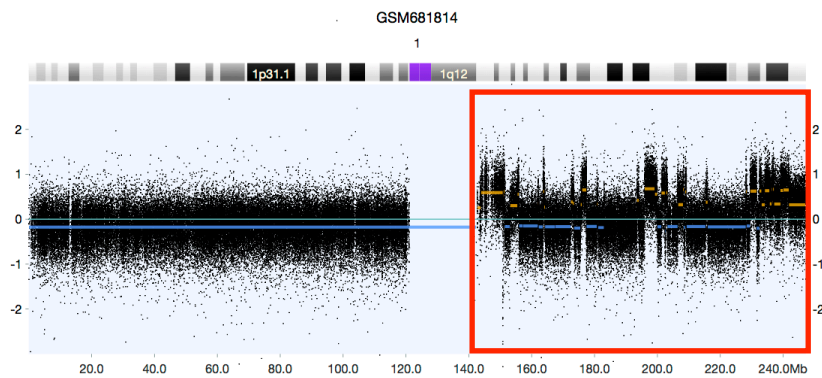


Log10(LR) = 57.7  
Switch times = 60  
Window size = 62.4 Mb

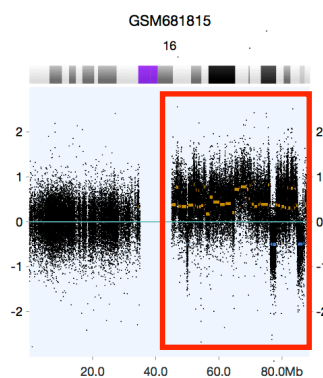


Log10(LR) = 26.7  
Switch times = 26  
Window size = 30 Mb

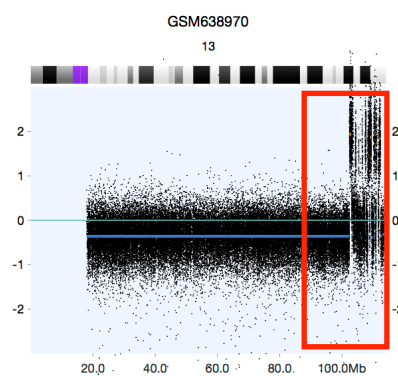
d



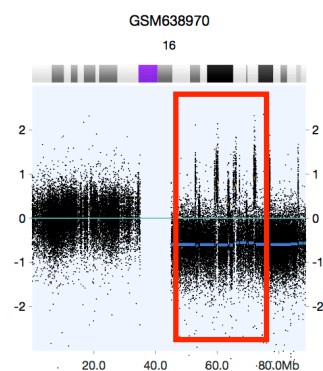
Log10(LR) = 50.8  
Switch times = 53  
Window size = 106.4 Mb



Log10(LR) = 98  
Switch times = 77  
Window size = 46.9 Mb

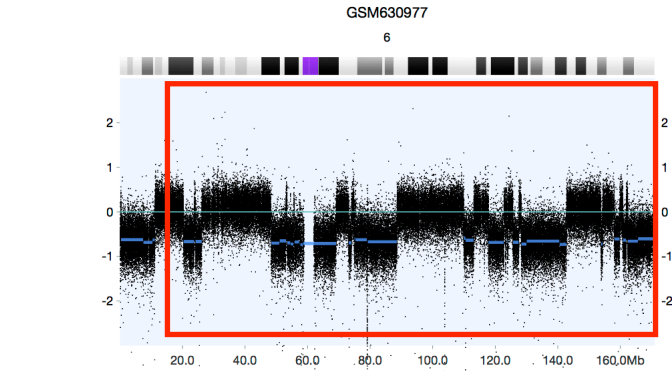


Log10(LR) = 42.3  
Switch times = 50  
Window size = 30 Mb

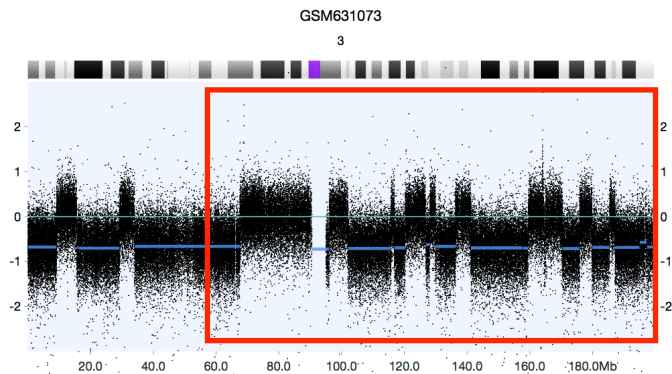


Log10(LR) = 34.7  
Switch times = 44  
Window size = 30 Mb

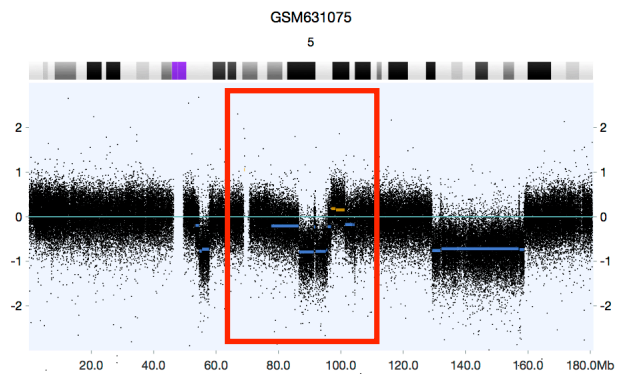
e



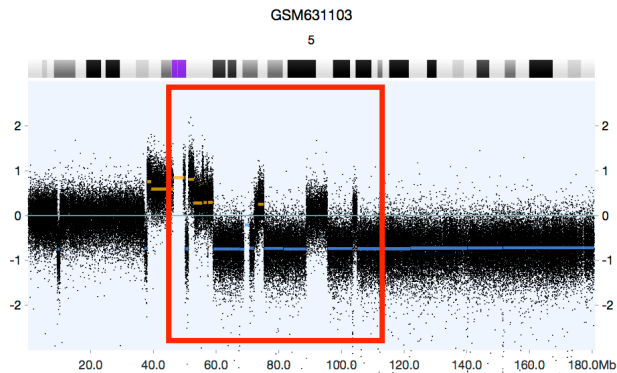
$\text{Log}_{10}(\text{LR}) = 31.8$   
 Switch times = 51  
 Window size = 146.3 Mb



$\text{Log}_{10}(\text{LR}) = 8.8$   
 Switch times = 24  
 Window size = 132.3 Mb

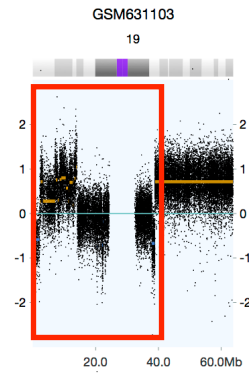


$\text{Log}_{10}(\text{LR}) = 10.7$   
 Switch times = 14  
 Window size = 40 Mb

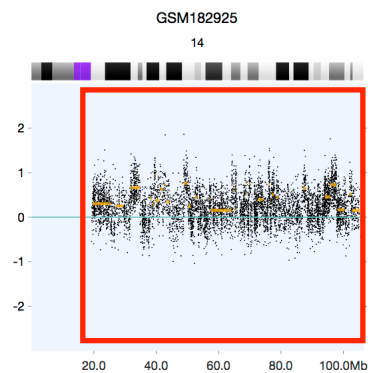


$\text{Log}_{10}(\text{LR}) = 18.4$   
 Switch times = 25  
 Window size = 62.4 Mb

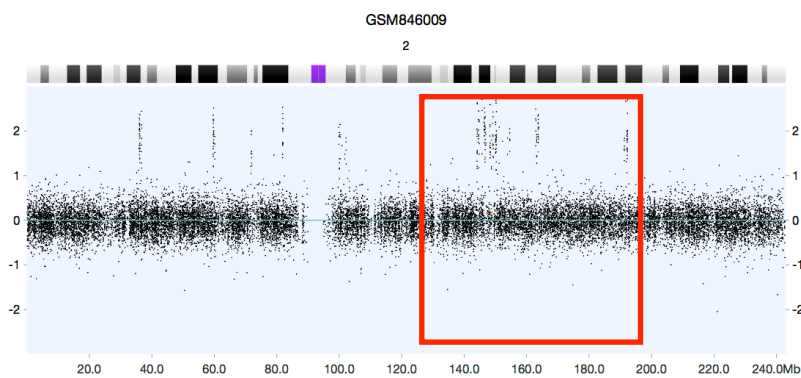
f



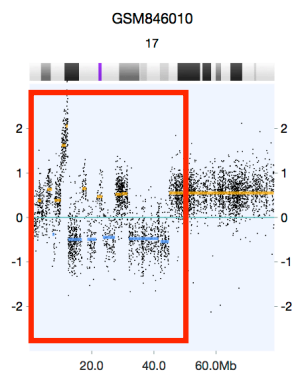
Log10(LR) = 28.7  
Switch times = 27  
Window size = 40 Mb



Log10(LR) = 57.7  
Switch times = 51  
Window size = 88.8 Mb

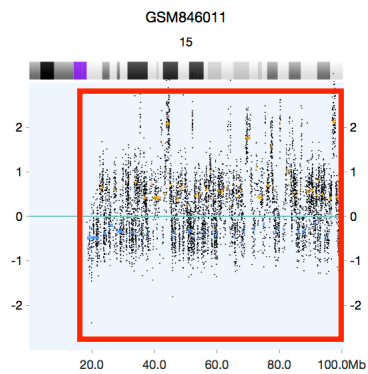


Log10(LR) = 34  
Switch times = 30  
Window size = 62.4 Mb

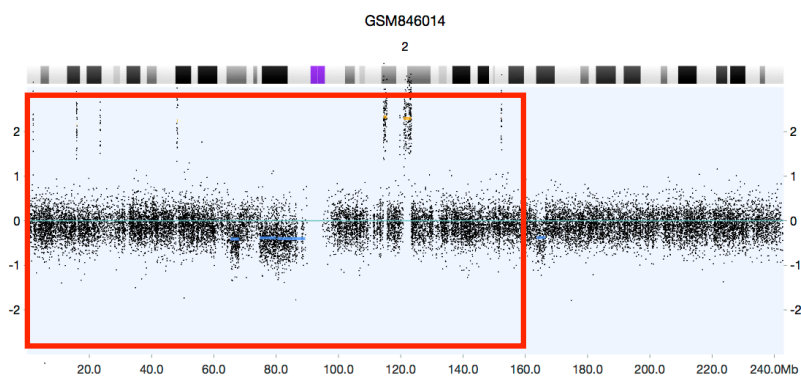


Log10(LR) = 16.4  
Switch times = 18  
Window size = 49.7 Mb

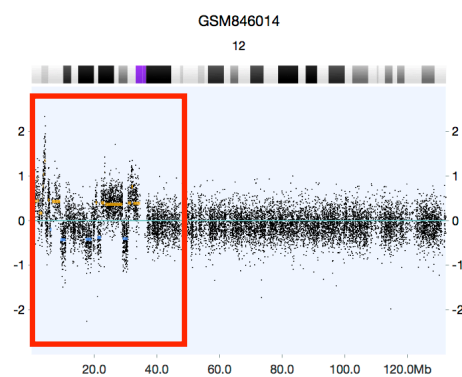
g



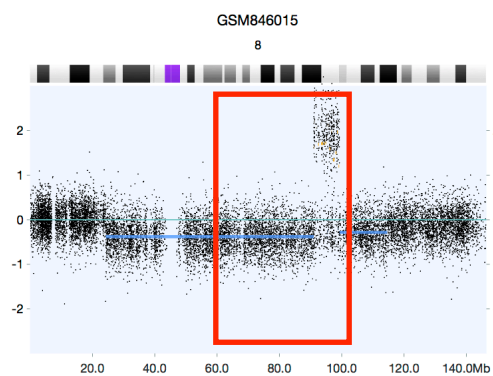
Log10(LR) = 240  
Switch times = 186  
Window size = 78.8 Mb



Log10(LR) = 11.1  
Switch times = 21  
Window size = 158.8 Mb

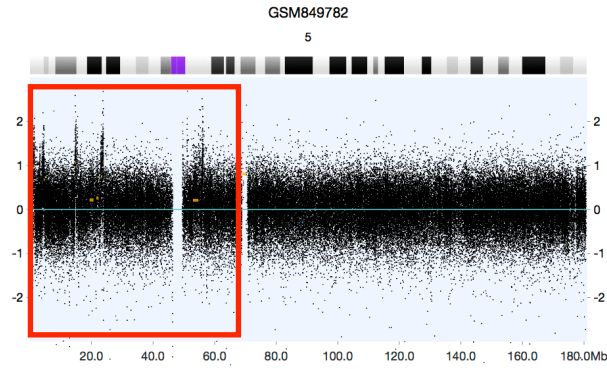


Log10(LR) = 25.1  
Switch times = 24  
Window size = 46.9 Mb

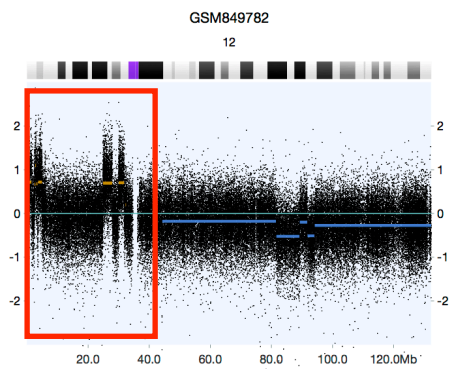


Log10(LR) = 48.5  
Switch times = 38  
Window size = 40 Mb

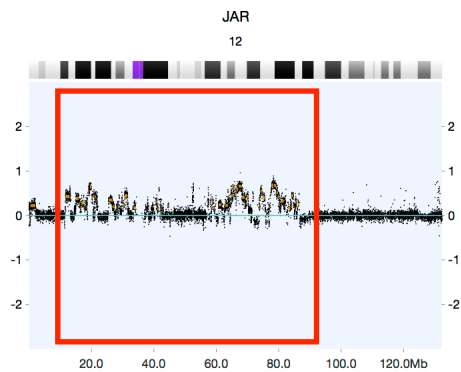
h



Log10(LR) = 30.4  
Switch times = 44  
Window size = 62.4 Mb



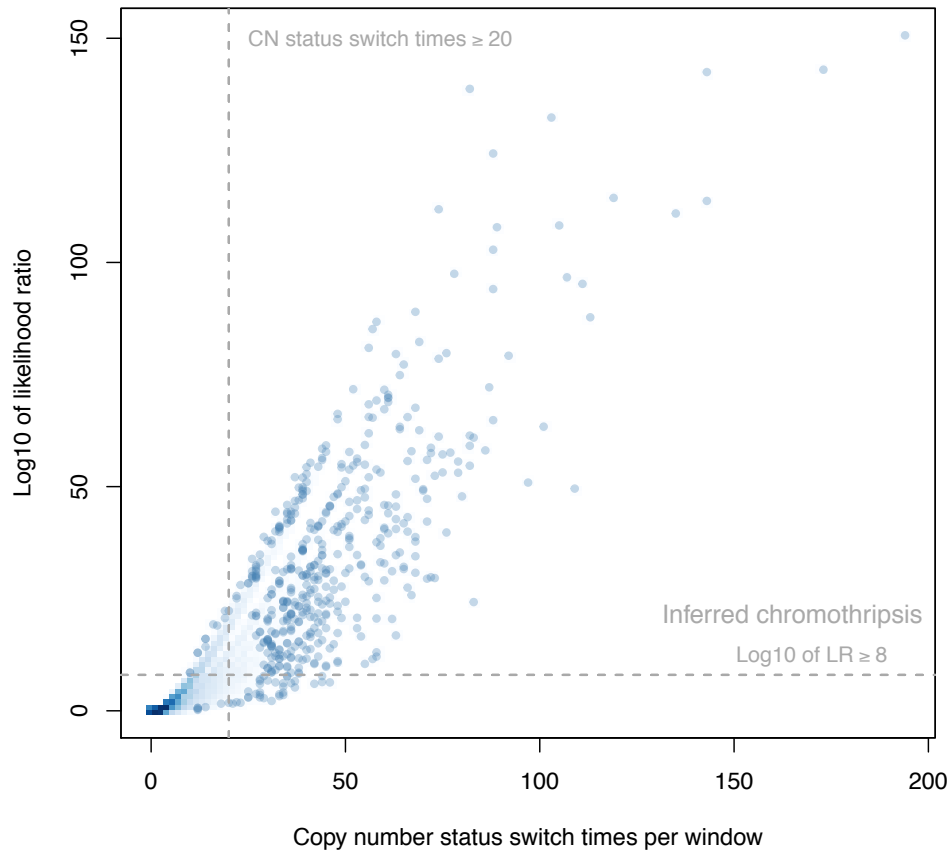
Log10(LR) = 12.6  
Switch times = 20  
Window size = 40 Mb



Log10(LR) = 107.8  
Switch times = 149  
Window size = 78.8 Mb

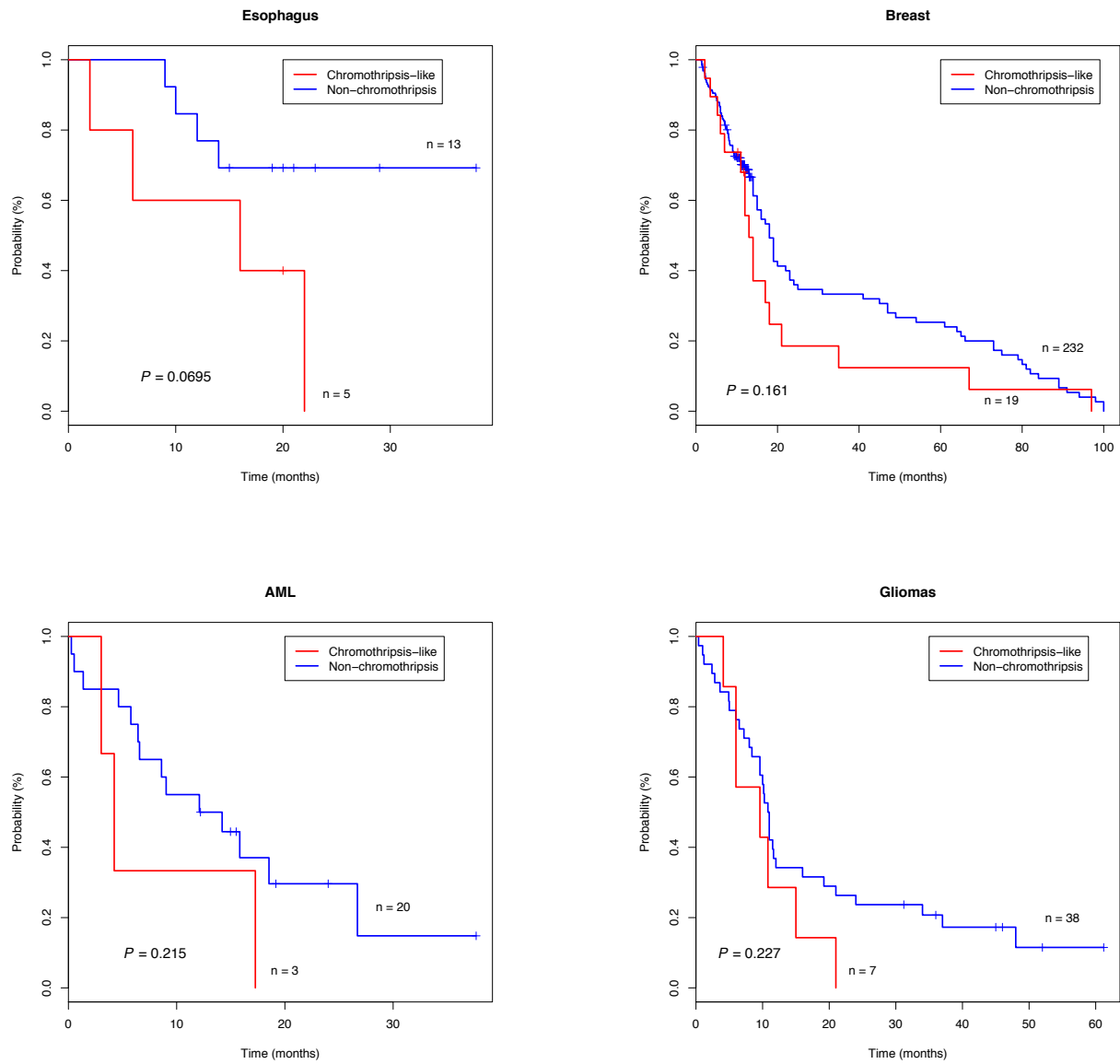
**Figure S2. The positive training set and CTLP detection algorithm performances.** The red rectangles are chromothripsis-like regions identified by scan-statistic. For each plot, the parameters and corresponding values are shown in orange boxes. The schema of the chromosome is the same as in Figure 2.



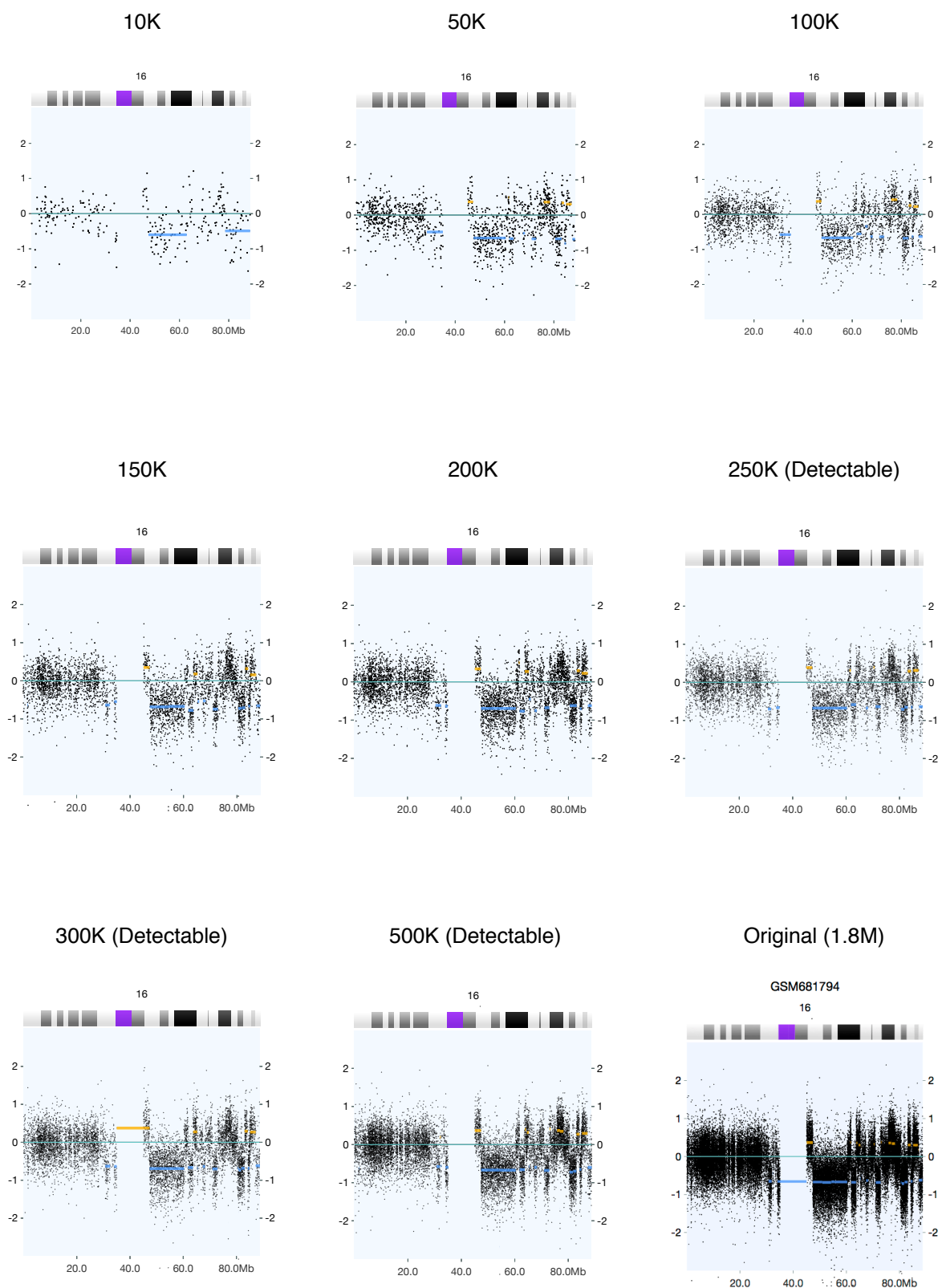


**Figure S3. Scatter plot of CTLP candidates.** For each chromosome of the input dataset, the window with the highest likelihood ratio was considered as a CTLP candidate. The selected thresholds are indicated with dashed lines. The candidates falling in the upper right area are inferred CTLP cases.

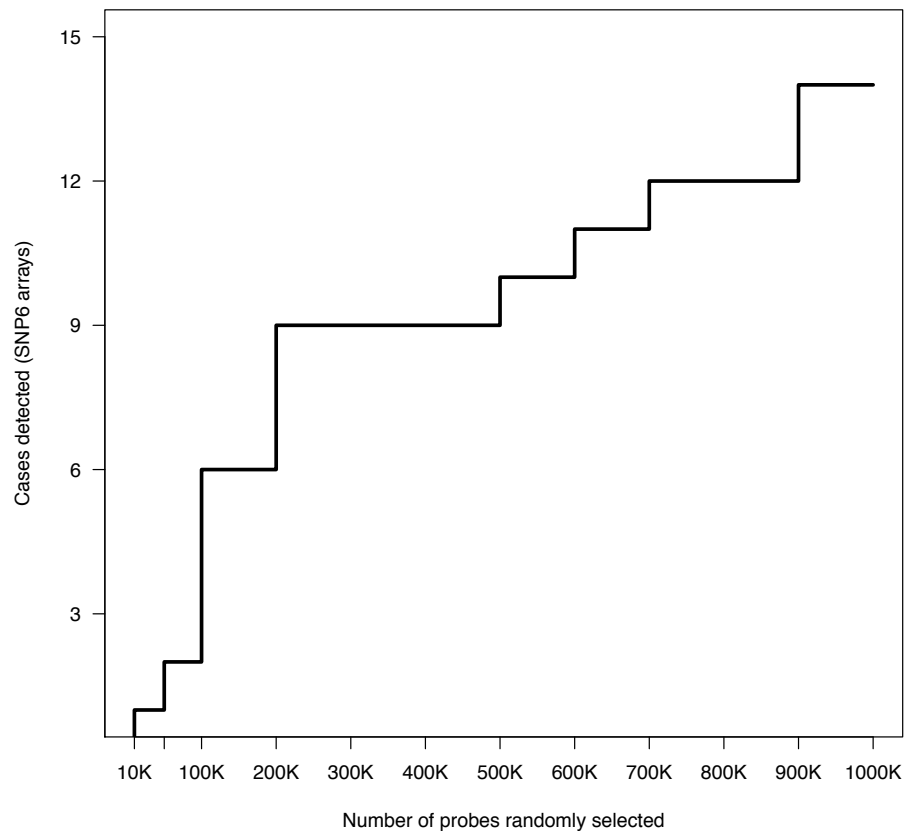




**Figure S4. Kaplan-Meier survival curves for CTLP versus non-CTLP cases in specific cancer types.**



**Figure S5. An example of the platform resolution based simulation from Affymetrix SNP6 array (1.8M).** Chromosome 16 of GSM681794 is a reported chromothripsis event in multiple myeloma. The number above each plot represents the number of probes randomly selected from the original probe set. In this sample, the CTLP pattern is detectable starting at 250k probes.



**Figure S6. CTLP detection sensitivity of simulated platform resolutions.** All the 15 chromothripsis chromosomes are from the positive training set analyzed using Affymetrix SNP6 platform.

**Table S1. Overview of input dataset**

Category	Array-level	Case-level
Total number	22347	18394
Series	402	397
Platform	190	185
Cancer type (ICD-O)	132	132
Cancer type (diagnostic group)	65	65
Source (primary)	19623	16309
Source (cell line)	2309	1714
Source (relapse)	75	75
Source (metastasis)	340	296

**Table S7. Demographic and clinicopathologic characteristics of input and CTLP samples**

Variable Name	Sample Number	Mean±SD/Median or %
<b>Input set</b>		
Male	1088	49.7%
Age(year)	2740	47.5 ± 25.7/55
<b>AJCC Stage</b>		
I	270	21.1%
II	331	25.9%
III	374	29.3%
IV	303	23.7%
<b>Grade</b>		
1	204	19.5%
2	435	41.7%
3	405	38.8%
Tumor recurrence	198	44.0%
Follow up (month)	1203	36.5 ± 34.6/26
<b>Event</b>		
Deceased	553	46.0%
Censored	650	54.0%
<b>CTLTP</b>		
Male	46	32.6%
Age(year)	259	54.1 ± 20.7/59
<b>AJCC Stage</b>		
I	14	14.3%
II	36	36.7%
III	34	34.7%
IV	14	14.3%
<b>Grade</b>		
1	16	14.4%
2	49	44.1%
3	46	41.5%
Tumor recurrence	18	47.4%
Follow up (month)	72	32.9 ± 35.2/17.6
<b>Event</b>		
Deceased	46	63.9%
Censored	26	36.1%

† All information is based on the available clinical data

**Table S8. Sizes of sliding windows for the scan-statistic based algorithm**

Size ID	Size (Mb)	Corresponding chromosome
1	247.249	Chromosome 1
2	242.951	Chromosome 2
3	199.501	Chromosome 3
4	191.273	Chromosome 4
5	180.857	Chromosome 5
6	170.899	Chromosome 6
7	158.821	Chromosome 7
8	146.274	Chromosome 8
9	140.273	Chromosome 9
10	135.374	Chromosome 10
11	134.452	Chromosome 11
12	132.349	Chromosome 12
13	114.142	Chromosome 13
14	106.368	Chromosome 14
15	100.338	Chromosome 15
16	88.827	Chromosome 16
17	78.774	Chromosome 17
18	76.117	Chromosome 18
19	63.811	Chromosome 19
20	62.435	Chromosome 20
21	46.944	Chromosome 21
22	49.691	Chromosome 22
23	40	NA
24	30	NA